An Ontology-Based Methodology for the Migration of Biomedical Terminologies to Electronic Health Records

Barry Smith, PhD[a,b], Werner Ceusters, MD[b]

[a]Department of Philosophy, University at Buffalo, Buffalo, NY 14260, USA
[b]Institute for Formal Ontology and Medical Information Science and European Centre for Ontological Research, Saarland University, 66041 Saarbrücken, Germany

Biomedical terminologies are focused on what is general, Electronic Health Records (EHRs) on what is particular, and it is commonly assumed that the step from the one to the other is unproblematic. We argue that this is not so, and that, if the EHR of the future is to fulfill its promise, then the foundations of both EHR architectures and biomedical terminologies need to be reconceived. We accordingly describe a new framework for the treatment of both generals and particulars in biomedical information systems that is designed: 1) to provide new opportunities for the sharing and management of data within and between healthcare institutions, 2) to facilitate interoperability among different terminology and record systems, and thereby 3) to allow new kinds of reasoning with biomedical data.

Keywords: terminologies, Relation Ontology, referent tracking, diagnostic decision support, SNOMED.

The General and the Particular

Much effort has been invested in recent years in the development of structured vocabularies of medical and biological terms. The resultant terminologies, such as SNOMED-CT[1], GALEN[2], ICD-10[3], and GO[4], consist overwhelmingly of general terms *A*, *B*, ... ('cell', 'tumor', 'postoophorectomy with pathological fracture') linked via relational assertions for example of the forms '*A is_a B*' or '*A part_of B*'.

The Electronic Health Record (EHR), in contrast, is a record of particular entities belonging to a wide variety of different general categories. It is a record of particular histories, of the events and processes described therein, of particular symptoms, disorders and associated pathological and non-pathological anatomical structures, of particular tests and particular measurements taken, and of much more.

For all this variety, however, most existing EHR architectures allow direct reference to just two kinds of particulars in reality:

(i) to *human beings* (patients, care-providers, family members) via proper names or via alphanumeric IDs,

(ii) to the *times* at which actions are performed or observations made.

For particulars in other categories existing EHRs provide, with few exceptions,[5] merely general codes. This limited repertoire of labels for particulars means that current EHRs have no adequate means to keep track of one and the same particular entity (for example a tumor, a breast implant, a shadow revealed on a succession of radiological images) over extended periods of time. When interpreting EHR data, it is thus difficult to distinguish clearly between multiple references to the same particular and multiple particulars of the same general kind.[5]

When the need arises to refer in different contexts to some single particular as it exists at different points in time, each such reference must at present be created anew, via some combination of general terms with designators for persons and times, for example in expressions like: *the fever of Patient McX first noted by Patient at t_1 and observed by Physician O'Z at t_2*. Unfortunately, the need to use such composites, and their syntactically *ad hoc* nature, creates logical obstacles to cross-identification of the corresponding entities as they occur in different contexts, and thus also to reasoning about these entities in software systems. This is so especially when we are concerned to keep track of how such entities develop over time, the capacity for which strikingly underdeveloped in existing biomedical information systems.

To resolve these problems, we have proposed a new type of EHR regime, in which, in the course of data entry, explicit alphanumerical IDs would be automatically assigned to each individual real-world entity at that point in time when it first becomes relevant to the treatment and care of a given patient.[5] Once assigned, such IDs would acquire the status of what we shall call IUIs, for *Instance Unique Identifiers*. IUIs would be assigned (in principle) not just to each particular tumor but also to each gland or duct in which a tumor is located, to each biopsy taken, to each associated radiological image, and indeed to instances in all the diagnostically salient categories recorded in the EHR.

Our proposal is that the physician or other specialist entering data about a given particular in the EHR should be provided with software tools that will constantly check the appropriate instance-tracking database in real time in order to establish as far as

possible automatically (with the aid of the SNOMED or other codes already entered) whether the particular in question is new to the system or whether it has already been allocated its own IUI. In the former case, a new IUI will be immediately created and subjected to coding in the usual way. In the latter case, the physician can simply add new information to the vector associated with the already existing IUI, using further codes as appropriate. The software will then make it easier to decide which codes to use at each successive stage, since the user will be presented immediately with the codes already entered in previous descriptions of the given particular.

We have outlined elsewhere both the benefits to diagnosis and treatment within a single healthcare institution which may derive from such an EHR architecture and also some of the practical problems involved in its realization.[5] Here we focus on those aspects of the referent tracking system which concern its interface with terminologies and coding systems.

A Map of Reality

Each IUI stored in the system (and thus each corresponding particular in reality) would be associated with a vector comprehending both relevant coding assignments and also cross-references to the measured values of medically salient attributes such as temperature or blood pressure, as well as gene expression and other bio-assay data. Such information would be annotated in its turn with time of entry, source, estimated evidence, access-rights, and so forth.

Importantly, the cross-referencing would be achieved by using the IUIs of those other particulars with which the entity under scrutiny is associated (for example the patient herself, her specific measurable attributes, family members, earlier events in her life history). Taken together, the vectors would thereby form a complex graph representing the relations between such particulars as they exist in reality. This graph would in addition contain large amounts of redundancy, of a sort which would enable cross-checking and thus error-detection in relation to the data entered in the associated EHRs.

Because instances of disorders would receive IUIs of their own, independently of any identifying reference to corresponding patients, the proposal would allow automatic compilation of pseudonymized data pertaining to specific kinds of disorders or to multiple disorders of specific kinds of patients. The ever-growing pool of vectors could further be managed in such a way that different kinds of associations between IUIs could be subject to different levels of encryption, thereby allowing new types of research collaboration based on the automatic exchange of different kinds of instance data.

In the ideal case, uniqueness of IUIs would be guaranteed by means of the same sorts of mechanisms as are currently used for maintaining the uniqueness of patient IDs or webpage addresses or of IDs for credit-card transactions. The vectors pertaining to particular identifiers might be stored locally or in some nationally or (ideally) internationally administered pool. In each case, our approach would facilitate the gathering of more adequate statistics on patient care and outcomes than are currently available.[6]

It will be important that referent tracking software have the facility to unravel ID assignments which have been discovered to be erroneous. When multiple IDs are assigned to what proves to be a single particular, it will normally be possible simply to merge the vectors associated with each separate IUI. In the (rarer) case where a single ID has been assigned to what turns out to be multiple particulars, software tools would need to be provided which would help the physician or coding specialist to decompose the corresponding vector in such a way that relevant segments come to be assigned to their associated particulars. In the same way, the system would need to be able to accommodate the sorts of corrections to the codes contained in specific vectors which become necessary when terminologies themselves change because of scientific advance.

All assignments and associated annotations would be preserved in their original forms for the medico-legal purposes of creating an audit trail. It is however a crucial feature of our proposal that it will allow active manipulation – or what we might think of as *tuning* – of the health record in real time on the part of the physician.[7] For the same tools which allow the correction of IUI mis-assignments would also provide the clinician with the opportunity to experiment with alternative IUI-assignments in support of reasoning about patients and their disorders. For example, successive clusters of symptoms of a given patient may be manifestations either of a single or of multiple disorders. Software could allow the physician to examine the consequences of rebundling vector annotations through association either with one or with a plurality of IUIs, and to allow statistical methods of pattern matching to compare the results of such rebundling as an aid to choice of diagnosis.

Moreover, by keeping track of the ways in which IUI-assignments are corrected in response to specific types of new information, the system could in principle learn to associate recurring patterns of correction – for example of the sort which arise in the early phases of diagnosis of degenerative diseases such as multiple sclerosis – with corresponding kinds of disorders.

The Terminology Problem

Even with an adequate system for tracking referents of the sort described, however, there is an obstacle to the effective migration of biomedical terminologies to the resulting EHR environment, which turns on the currently predominant treatment of the terms and relations in such terminologies.

In the development of almost all extant terminologies, little consideration was given to the need for a clear link, or bridge, between terms in terminologies and instances in reality. Rather, the relations between terms were (and often still are) conceived primarily in the ways in which linguists or dictionary-makers might conceive them, which is to say: as reflecting merely certain relations between *meanings of words*.

Consider for example the definition of the *is_a* relation provided by the Semantic Network (SN) of the Unified Medical Language System (UMLS)[8]:

If one item 'isa' another item then the first item is more specific in meaning than the second item.

The nodes of the SN – called, variously, 'concepts' or 'items' or 'entities' or 'Semantic Types' – would seem in light of this definition to be precisely *meanings*. However, when we examine the bulk of the SN's relations between such nodes, we see that they are defined in such a way as to require instead a conception of these nodes as *entities in reality*. *Part_of*, for example, SN defines as: *composes, with one or more other physical units* .... The relation *contains* is defined as: *holds or is the receptacle for fluids or other substances*. *Co-occurs_with* as: *occurs at the same time as, together with, or jointly*. Meanings, clearly, cannot serve as the relata of such relations. Our suspicion is that there is no answer to the question what the nodes of the SN are which would make all its relational assertions come out true simultaneously.[7,9]

The SN is of course not designed to be used as a terminology in healthcare records. But the problem of polysemy of the term 'concept' applies to almost all the terminologies included in the UMLS Metathesaurus, where relations like *synonymous_with* or *narrower_in_meaning_than* or *associated_with* or *conceptually_related_to* are used side –by side with relations like *treats* or *causes* or *is_finding_site_of*.

Because the relations which structure these terminologies were introduced in informal and inconsistent ways, the logical interconnections between the corresponding assertions are left unclear. This goes far, we believe, to explain familiar errors in coding and documentation.[10] The same shortfall has also served to block those kinds of logical reasoning within and between terminologies and EHRs which would be possible given clear and consistent definitions.[11]

A New Regime of Definitions

How, then, are we to reconstruct biomedical terminologies in such a way that they will both (1) allow the provision of clear definitions of relations such as *is_a* and *part_of*, and at the same time (2) facilitate application to corresponding instances and times in reality and thus to the EHR? The answer proposed in [12] consists in a treatment of the relations used in biomedical ontologies as linking, not concepts or meanings, but rather *entities in reality* at the level of both universals and their instances. (The term 'universal' here refers to those invariants in reality, sometimes called 'classes' or 'types', in virtue of which we are able to describe multiple particulars by using one and the same general term. It is such invariants which make possible *inter alia* the application of standardized therapies to multiple instances of the same disorder (universal) in different patients.) The method thus explicitly accommodates, in addition to the familiar relations involving universals or types (marked in *italics* in what follows), also relations involving their underlying instances (marked in **bold**).

The Relation Ontology which results from the application of this method has been incorporated into the ontology library maintained by the Open Biomedical Ontologies Consortium,[13] and it is being used by curators of the Foundational Model of Anatomy,[14] of GO[4], and of the ChEBI chemical entities vocabulary[13] as part of their work on ontology integration designed to support more powerful cross-domain reasoning and data annotation. The Relation Ontology is designed also to support more reliable curation of ontologies, since its definitions have been formulated in such a way as to provide an optimal combination of understandability to curators of ontologies with the sort of formal rigor needed to support logic-based reasoning.

We here provide examples of definitions from the Relation Ontology pertaining to *continuant entities* such as lungs, diseases, tumors, fractures, entities that endure through time while undergoing changes of various sorts (as contrasted with *occurrent* entities, which unfold themselves through time in successive phases.[12,15]). We use variables c, d, ..., C, D, ... to range over continuant particulars and continuant universals, respectively. Because the former can instantiate different universals and include different instance-level parts at different times (consider, for example, a carcinoma, or a fetus, in its successive stages of development), our definitions require also variables t_1, t_2, ... to range over instants of time (assumed to be linearly ordered by the relation **earlier_than**). They require also certain primitive (undefined) relations involving continuant entities on the instance level, for example:

c **instance_of** *C* **at** *t* (particular *c* instantiates universal *C* at time *t*);

c **part_of** *d* **at** *t* (particular *c* is an instance-level part of particular *d* at time *t*);

c **located_in** *d* **at** *t* (the spatial region occupied by *c* is an instance-level part of the spatial region occupied by *d* at time *t*).[16]

Corresponding formal definitions of relations between continuant universals then read as follows:

C is_a D =def. for all *c*, *t*, if *c* **instance_of** *C* **at** *t* then *c* **instance_of** *D* **at** *t*;

C part_of D =def. for all *c*, *t*, if *c* **instance of** *C* **at** *t* then there is some *d* such that: *d* **instance_of** *D* **at** *t* and *c* **part_of** *d* **at** *t*;

C located_in D =def. for all *c*, *t*, if *c* **instance_of** *C* **at** *t* then there is some *d* such that: *d* **instance_of** *D* **at** *t* and *c* **located_in** *d* **at** *t*;

C transformation_of D =def. for all *c*, *t*, if *c* **instance_of** *C* **at** *t*, then there is some t_1 such that: *c* **instance_of** *D* **at** t_1 and t_1 **earlier than** *t*.

The relation *transformation_of* serves to represent the phenomena of growth, development and pathological change. Where this relation obtains between two universals *C* and *D* (for example *adult* and *child*), then every instance of the former was at some prior stage an instance of the latter (so that we have some single continuant particular which instantiates distinct universals at different times in virtue of phenotypic changes).

Note how the definitions listed ensure that the corresponding relational assertions on the level of universals have embedded within them an automatic reference to the corresponding instances and times. This is achieved characteristically via an *all–some* structure,[17,18] as for example in the definition of *part_of*, above, where we have: universal *C part_of* universal *D* when **all** instances of *C* have **some** instance of *D* as part.

Most existing biomedical terminologies employ an informal treatment of relations that fails to distinguish clearly between relations on the two levels of instances and universals. At the same time they commonly fail also to distinguish *all-some* (*human has_part lung*) from *some-some* relations (*human has_part testis*), in ways which thwart the drawing of reliable inferences.[17,18] With a properly formal treatment of relations, in contrast, it becomes possible to reason across data deriving from different sources, secure in the knowledge that our inferences will track the underlying reality. Thus, given *all-some* relations R_1 and R_2, if we know from ontology$_{(1)}$ that *A* R_1 *B* and from ontology$_{(2)}$ that *B* R_2 *C*; and if we have instance-data concerning some *A* from some referent-tracking data-source, then we know (1) that this *A* stands in the instance-level counterpart of R_1 to some *B*, and also (2) that whichever *B* this is, it stands in its turn in the instance-level counterpart of R_2 to some instance of *C*. Our software can then be made to search for IUIs for the corresponding instances of *B* and *C* within all vectors including cross-references to the IUI of our initial instance of *A*.

In the EHR domain, reasoning of this sort might be used to integrate data in an EHR with information contained in a system such as the TNM (for: Tumour, Node, Metastasis) classification of cancer stages. Here 'T1' designates a stage where the tumor has invaded the submucosa, 'N1' a stage with one to four lymph nodes, 'M1' a stage where a metastasis is present in a non-contiguous part of the body, and so forth. In a vector describing the patient, there might figure IUIs that refer to one or several metastases, to processes of growth, and so on, and whose vectors will incorporate descriptors for the corresponding TNM configurations. Using the machinery of the Relation Ontology, we can then extract from TNM a system of assertions to the effect, for example, that the universal T2N1M1 *carcinoma in colon* stands in a *transformation_of* relation to either a T1N1M1 or a T2N0M1 structure.[19] Such assertions can then be used by our software to infer testable hypotheses about the type of carcinoma development under scrutiny.

A New Regime for Clinical Coding

Suppose, against this background, that we wish to use the resources provided by a terminology such as SNOMED in order to draw inferences concerning a specific patient suffering from recurring breast cancer. History-taking involves finding ways of referring to instances such as: this present incidence of breast cancer, earlier incidences, present tumors, earlier (distinct) tumors, processes of mastectomy, and so on. How, then, can we use standard SNOMED-coded EHR data in order to keep track of these multiple particulars at successive points in time if (as in present EHR architectures) we have unique IDs only for patients and for time-points?

One answer[20] holds that we can provide some of what we need to achieve this end through inferences from statements at the general level. If, for example, a given SNOMED term is used at t_1 to describe 'something a physician observed' and at t_2 the same general term is used again by the same physician, and if t_2 is 'close in time' to t_1, then it can be inferred that the physician referred to the same 'something' on two successive occasions.

We find this idea practically implausible. Indeed, while we fully support the idea of inferencing from SNOMED-coded data, we believe that such inferencing will become feasible only in the presence

of large amounts of instance data accumulated along the lines described above. This is because, for the reasons explained, many of SNOMED's existing relational assertions cannot be used to infer further information about particular instances, and SNOMED currently offers no way to tell which of its assertions do support inferences of this sort. Thus its relational organization is still best conceived as a convenient mechanism for browsing through the terminology in order to find better descriptors, not as a representation of how the corresponding instances are related together in reality. When the paradigm here advanced has been in use for some time, however, then the accumulated instance-data could be exploited *post hoc* to correct SNOMED's treatment of relations in such a way that it would, by degrees, be in a position to support such inferences in a reliable fashion.

In this way our methodology can also lead to improvements in terminologies. It can also further the goal of interoperability between terminologies and other systems for recording biomedical data. For our paradigm would allow the simultaneous use of a variety of different coding systems within a single record. This use of multiple codes would then yield in automatic fashion an ever-growing network of associations between the terms in the separate coding systems – reflecting their use in annotating common particulars – in a process which would eventually supplant current efforts to create mappings between such systems.

Conclusion

There is however one remaining obstacle to the use of formal definitions to support reasoning with instance-level data, which turns on the fact that such reasoning is expensive in computational resources. As is recognized in Description Logic circles, however, the right approach is to concentrate first on those problems that allow tractable reasoning and to focus on the hard cases later. We are currently testing a prototype reasoner which can help us evaluate the potential of the method along these lines. Given assertions of specific relations between instances of given types, our prototype calculates, on the basis of definitions in the OBO Relation Ontology, an exhaustive list of all further relations which can hold between instances of the types in question. It thereby becomes possible to transform *reasoning* with instance data into *search* across the corresponding relation space, which entails far fewer demands on computational resources. A framework is hereby being built which can, we believe, help in bringing together in dynamic fashion the distinct ways of treating data that have evolved in the worlds of clinical records and of medical terminology.

References
1. http://www.snomed.org.
2. Rogers J, Rector A. The GALEN ontology. *MIE* 1996;:174-178.
3. http://www.who.int/classifications/icd/en.
4. The Gene Ontology: http://www.geneontology.org.
5. Ceusters W, Smith B. Tracking referents in electronic health records. *MIE* 2005. In press.
6. Ceusters W, Smith B. Strategies for referent tracking in electronic health records. J Biomed Inform, http://ontology.buffalo.edu/medo/Tracking-referents.pdf.
7. Smith B. New desiderata for biomedical terminologies. J Biomed Inform (submitted).
8. http://semanticnetwork.nlm.nih.gov.
9. Smith B. Beyond concepts: Ontology as reality representation. Formal Ontology and Information Systems (FOIS 2004);:73-84.
10. Farhan J, Al-Jummaa S, Al-Rajhi A, Al-Rayes H, Al-Nasser A. Documentation and coding of medical records in a tertiary care center: a pilot study. Ann Saudi Med. 2005 Jan-Feb;25(1):46-9.
11 Rector AL. Clinical terminology: why is it so hard? Methods Inf Med. 1999;38(4-5):239-52.
12. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A, Rosse C. Relations in biomedical ontologies. Genome Biology 2005, 6:R46.
13. http://obo.sourceforge.net.
14. Rosse C, Mejino JLV. A reference ontology for bioinformatics: The Foundational Model of Anatomy. J Biomed Inform 2004;36:478-500.
15. Rosse C, Kumar A, Mejino JLV, Cook DL, Detwiler DL, Smith B. A strategy for improving and integrating biomedical ontologies. *AMIA* 2004 (in this volume).
16. Donnelly M: Layered mereotopology. *Proc IJCAI* 2003;:1269-1274.
17. Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. Proc. Medinfo 2004; 444-448.
18. Donnelly M, Bittner T, Rosse C. A formal theory for spatial reasoning in biomedical ontologies. *Artificial Intelligence in Medicine*, in press.
19. Kumar A, Yip YL, Smith B, Grenon P. Bridging the gap between medical and bioinformatics using ontological principles: A colon carcinoma case study, Computers in Biology and Medicine, in press.
20. http://www.cs.man.ac.uk/mig/projects/current/clef.